Elsevier Editorial System(tm) for Learning and Instruction

Manuscript Draft

Manuscript Number: JLI-D-15-00391R3

Title: An Oracy Assessment Toolkit: linking research and development in the assessment of students' spoken language skills at age 11-12

Article Type: SI: Classroom discourse

Keywords: oracy
spoken language
talk
assessment

Corresponding Author: Mr. Paul Warwick, MEd

Corresponding Author's Institution: University of Cambridge

First Author: Neil Mercer

Order of Authors: Neil Mercer; Paul Warwick, MEd; Ayesha Ahmed

Abstract: This article describes the development of a set of research-informed resources for assessing the spoken language skills (oracy) of students aged 11-12. The Cambridge Oracy Assessment Toolkit includes assessment tasks and procedures for use by teachers, together with a unique Skills Framework for identifying the range of skills involved in using talk in any specific social situation. As we explain, no comparable, 'teacher-friendly' instrument of this kind exists. Underpinning its development is the argument that teaching children how to use their first or main language effectively across a range of social contexts should be given higher priority in educational policy and school practice, and that the development of robust, practicable ways of assessing oracy will help to achieve that goal. We explain how the Toolkit has been developed and validated with children and teachers in English secondary schools, and discuss its strengths and limitations.

UNIVERSITY OF CAMBRIDGE
Faculty of Education

September 2016

To the Editors

We wish to thank our reviewers and the editors of the special edition on Classroom Discourse for the hard work that they have put in to reviewing our paper 'An Oracy Assessment Toolkit: linking research and development in the assessment of students' spoken language skills at age 11-12'. We now feel that we have a paper ready for publication that has been forged in response to rigorous constructive criticism.

Regards,

Paul Warwick
For Neil Mercer, Paul Warwick and Ayesha Ahmed

Head of Faculty: Professor Geoff Hayward Secretary to the Faculty: Kate Allen
Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge CB2 8PQ, UK
Tel: +44 (0) 767600 Fax: +44 (0) 767602 Email: reception@educ.cam.ac.uk www.educ.cam.ac.uk

Responses to editor and reviewers – June ‘16

We have inserted the following paragraph in Section 1.4:

Most importantly, this approach does not deny that different tasks, curriculum subject areas and genres have distinctive features in terms of the development of oracy skills, and that these can be highly specialised, applying only to a single genre for example. However, the generic skills-based framework developed and used within the project (Figure 1, below) has a very specific function. It provides an over-arching framework of *generic* skills - categorised as physical, linguistic, cognitive and social – from which relevant skills can be *selected* for assessment as relevant to a given task. So, ‘building on the views of others’ might be highly pertinent in an assessment of group talk, but not necessarily in public speaking; yet ‘fluency and pace of speech’ might pertain to a drama performance, a presentation and so on. The important thing is the selection of skills for assessment and their association with particular tasks, enabling the teacher to build a profile over time. Thus, we would not suggest that this framework is completely comprehensive for all language use in all contexts, curriculum areas or subjects. Nor do we suggest that the *whole* framework of skills is relevant to every context. However, we demonstrate how several assessments, used across time and in a range of contexts, can build a generic oracy profile for a student. Further, we demonstrate that this is only really possible if teachers have an overall framework of broad oracy skills as a template for their various assessments.

*Title page (including all author details and affiliations)

An Oracy Assessment Toolkit: linking research and development in the assessment of students' spoken language skills at age 11-12

Neil Mercer, Paul Warwick* and Ayesha Ahmed
Faculty of Education, University of Cambridge

*Corresponding author: ptw21@cam.ac.uk; Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge, CB2 8PQ

Keywords: oracy; spoken language; talk; assessment.

***Highlights (for review)**

- Evaluation of research-informed resources for assessing spoken language skills.
- ‘Oracy Assessment Toolkit’ has assessment tasks and a unique Skills Framework.
- Review of development, validation, strengths & limitations of Toolkit.

1. Introduction

If it is accepted that schools should be helping students to develop effective talk skills, then teachers need practical ways of monitoring and assessing the oracy skills of their students in a classroom setting. Useful schemes for assessing children's language development are available, but surprisingly no suitable assessment instruments seem to exist for children aged 11-12 (the start of secondary education in the UK and many other countries). Moreover, no clear skills framework exists for identifying the different aspects of spoken language use that young people need for the range of communication situations they will encounter. Such tools would help teachers to plan how to use classroom discourse to enable their students to become more metacognitively aware, and more skilled speakers and listeners.

1.1 The importance of spoken language education and assessment

In recent years, researchers in developmental psychology, linguistics and education have emphasised the importance of talk for stimulating children's cognitive development, and its use as both a cognitive and social tool for learning and social engagement (see for example van Oers, Elbers, van der Veer & Wardekker, 2008; Whitebread, Mercer, Howe and Tolmie, 2013). In doing so, they follow Vygotsky (1962) who gave the acquisition of language a crucial place in his model of cognitive development. As Vass and Littleton have put it, 'interpsychological thinking is a prerequisite for intrapsychological thinking: it is through speech and action with others that we learn to reason and gain individual consciousness.' (2010, p. 107). Research in neuroscience and evolutionary psychology supports the view that language has evolved as an integrated component of human cognition, rather than as a separate and distinct capacity (Goswami, 2009; Mercer, 2008, 2013; cf. Pinker, 2007).

Like many capacities, language development is affected by the quality of experience. Our view is that oracy (like literacy) consists of a range of diverse skills, which may develop/be learned to different extents; and for many children only some skills may have been modelled and encouraged (by other people) in their out-of-school experience. By the time they reach secondary school, some children may have learned how to carry on informal conversations and to engage in lively banter with their peers, but not have developed the ability to speak confidently to a public audience or engage in a reasoned debate. Some may have developed much larger vocabularies than others. So, individual children's experience may generate very different oracy profiles. Such diversity may affect children's ability to participate in the language-based process of school education. Research has shown that the amount and quality of pre-school children's conversations in the home are good predictors of educational attainment in secondary school (Goswami & Bryant, 2007; Hart & Risley, 1995). A systematic review of research (Howe & Abedin, 2013) has found positive associations between student learning and the use of extended and cumulative responses in group interactions; and such responses often result from specific teaching about how to use talk effectively to engage in reasoned discussions (Dawes, 2008). Overall, this encourages the view that the extent to which schools give

direct attention to oracy can influence students' learning and cognitive development through building their ability to use language effectively across a range of contexts. Further, it seems that the development of certain ways of using language can influence students' future social mobility; indeed, there are studies that are starting to show that there is a strong relation between oral communicative competence and social acceptance and status (van der Wilt, F., van Kruistum, C., van der Veen, C., & van Oers, B., in press).

There is also growing recognition amongst those outside formal education of the importance of all young people learning to use talk effectively for social and democratic engagement, and in work-related activities. An expert report on skills for employability commissioned by the London Chamber of Commerce stated: 'Softer skills, such as team working and communication, are an important aspect of an individual's employability, and they will be in higher demand as we move towards a more knowledge-intensive economy.' (Wright, Brinkley & Clayton, 2010, p. 8). This is why we believe that oracy needs to be assessed, and taught according to need. For many children, the only hope of developing a full repertoire of oracy skills is if oracy is given the same kind of attention in school that has traditionally been given to literacy.

The various issues discussed above persuaded us that the development of oracy deserves more attention in schools; and that giving it that attention would be assisted by the development of a valid, reliable but practical way for teachers to monitor and assess the spoken language skills of their students. In partnership with *School 21* (a school in London which has put oracy at the core of its curriculum[1]) we bid successfully to the Educational Endowment Foundation for funds to develop an 'Oracy Assessment Toolkit'. In summary, our motivation for developing the Toolkit was thus based on the following concerns:

(a) the development of students' spoken language skills (oracy) is as important for their future lives as the development of their literacy and numeracy;
(b) they need oracy skills to participate effectively in classroom life and in wider society;
(c) like literacy and numeracy, oracy can be taught and assessed;
(d) oracy is more likely to be recognised as an important part of the school curriculum if it can be assessed;
(e) teachers need to assess the strengths and weaknesses of their students' spoken language skills if they are to provide suitable guidance and instruction, and they need to be able to assess the effects of their teaching on students' skills;
(f) there are currently no 'teacher-friendly' tools available for assessing children's oracy at the age at which they commonly begin secondary school (which is normally 11 years old in the UK).

1.2 Defining oracy

[1] School 21 is a 'free school', meaning that its founders have gained direct funding from the national government to establish a school which is not constrained by the National Curriculum for England and Wales. See http://school21.org.uk.

Andrew Wilkinson introduced the term 'oracy' to refer to 'the ability to use the oral skills of speaking and listening' (Wilkinson, 1965, p.13).It is our view that the introduction and use of a concept to describe children's overall ability to use spoken language is extremely valuable and justifiable, especially in relation to educational research, policy and practice. The same applies of course to the use of 'literacy' and 'numeracy': Wilkinson (1965) coined the term 'oracy' in order to try to give spoken language skills comparable status, and to help resist narrow, back-to-basics conceptions of school curricula which typically do not accord talk a similar status to reading and writing. Some researchers, policy makers and practitioners readily adopted Wilkinson's term and definition, as in the UK's National Oracy Project (Norman, 1992). However, other terms such as 'communication skills' and 'speaking and listening' have tended to be used more widely in the English speaking world (DfES, 2003). Alexander has argued that such terms "have become devalued by casual use" (Alexander, 2012, p.2) and thus 'oracy' represents the best way to refer to "children's capacity to use speech to express their thoughts and communicate with others, in education and in life" (ibid, p.10). Agreeing with those sentiments, in this paper 'oracy' is used to refer to the development of young people's skills in using their first language, or the official/educational language of their country, to communicate across a range of social settings.

1.3 Challenges in assessing oracy

Howe (1991) described three main challenges for the assessment of oracy: the fact that spoken language is ephemeral; the restriction on the number of students that can be assessed at a time; and the context specificity of speech acts. He echoed Barnes'(1980) argument that to assess fairly we need a wide range of contexts in which to gather evidence. In such contexts, Cinamon & Elding (1998, p. 220) define progress as "gaining increasing control over…language to a wider range of audiences, for a greater variety of purposes and in different settings".

Teachers commonly feel less confident about what constitute oracy skills in comparison with literacy skills. For example, Haig & Rochecouste (2005) interviewed Australian teachers in 13 secondary schools, concluding that they had a narrow concept of competence in oracy, mainly identifying it with the ability to make formal public presentations. Talk-based activities, such as group work, were considered to be 'peripheral to performance' (ibid, p. 218) Overall, these teachers felt that they 'do not have the skills to assess oral language' (ibid, p. 212).

An additional challenge in oracy assessment is that, in many situations, talk involves the integrated activities of two or more people; how can individual performance be isolated? For assessing talk in group tasks, Wilson, Neja, Scalise, Templin, Wiliam & Torres Irribarra (2012) suggest that each individual's performance should be based on the aggregate of their performances over many groups and over multiple contexts, as well as using feedback from all group members about each individual's contribution. However, this would be impractical in 'normal' classroom settings. The PISA 2015 assessments (OECD, 2014) recognise the importance of skills in collaborative thinking, though the

OECD has chosen to assess collaborative problem-solving skills individually, with a computer agent acting as the other group member. This avoids the problems of assessing an individual within a group, but removes the normal social features of a group of real students.

1.4 How has oracy been assessed?

Assessments of children's talk skills have commonly been concerned with competence in a specific speech genre, such as taking part in a debate, presenting a prepared monologue in public (Monroe, 2009), or engaging in collaborative problem solving (Mercer & Littleton, 2007). Such genre-based approaches only provide a limited picture of a child's overall competence. Moreover, our contact with school head-teachers and English specialists encouraged us to believe that they would ideally like to obtain a generic assessment in oracy for each child (comparisons with 'reading age' had been made in such discussions). Our aim was to enable this possibility (or a rather more finessed approach) by devising a set of specific, situational tasks to provide a profile of a child's oracy skills across a range of situations. The broad approach adopted was to first devise an over-arching oracy skills framework (see Section 2.2.1 and task examples below) and then to consider possible test tasks in the light of this framework. The framework could be applied to any suitable talk task so that the core oracy skills within it could be identified for assessment and so that a more holistic assessment of could be arrived at across tasks.

Most importantly, this approach does not deny that different tasks, curriculum subject areas and genres have distinctive features in terms of the development of oracy skills, and that these can be highly specialised, applying only to a single genre for example. However, the generic skills-based framework developed and used within the project (Figure 1, below) has a very specific function. It provides an over-arching framework of *generic* skills - categorised as physical, linguistic, cognitive and social – from which relevant skills can be *selected* for assessment as relevant to a given task. So, 'building on the views of others' might be highly pertinent in an assessment of group talk, but not necessarily in public speaking; yet 'fluency and pace of speech' might pertain to a drama performance, a presentation and so on. The important thing is the selection of skills for assessment and their association with particular tasks, enabling the teacher to build a profile over time. Thus, we would not suggest that this framework is completely comprehensive for all language use in all contexts, curriculum areas or subjects. Nor do we suggest that the *whole* framework of skills is relevant to every context. However, we demonstrate how several assessments, used across time and in a range of contexts, can build a generic oracy profile for a student. Further, we demonstrate that this is only really possible if teachers have an overall framework of broad oracy skills as a template for their various assessments.

To consider previous assessments of oracy, one of the first attempts to make a holistic assessment of children's oracy was made in the UK by The Assessment of Performance Unit. Their survey (APU, 1988) monitored thousands of students aged 11 and 15, and included tasks designed to assess their oracy skills. Tasks included presentations and

paired problem solving activities. Students were assessed by trained assessors. Their main conclusions were that it is feasible to monitor speaking and listening performance on a national scale; that marker reliability was satisfactory; and that the assessment materials had communicative validity. They noted that 'almost all 11 year olds can modify their speaking strategies appropriately in accordance with the demands of different tasks and different audiences' (APU, 1988, p. 64). In a similar study in the Netherlands, the oracy skills of two hundred 10 to 12 year old students were measured by the research team (van den Bergh, 1987). Six tasks were constructed and the study concluded that the assessment of oracy is feasible for this age group, with only 13% of the students failing or responding at a 'doubtful' level. Maybin (1988) criticised the APU's assessments because they were based on performances made outside normal contexts and were based on an individualistic, non-interactive model of language use. The problems of reducing a social phenomenon to a series of assessment tasks are made clear in Maybin's critique.

Recently, test developers have designed more interactive tasks. Latham (2005) created a Speaking and Listening Profile to help teachers to use the Speaking, Listening and Learning materials that supported the English National Curriculum; and diagnostic in-school assessment schemes for teachers in the UK have been devised by the Qualifications and Curriculum Development Agency (QCDA, 2008; 2010). These contained four assessment foci for speaking and listening (talking to others, talking with others, talking within role-play and drama, and talking about talk) and four strands of relevant oracy skills (listening and responding; speaking and presenting; group discussion and interaction; drama, role play and performance). However, from summer 2014 a speaking and listening component will no longer count towards the final General Certificate of Secondary Education (GSCE) grade for examinations in English; this followed 'concerns about the effectiveness of the moderation of controlled assessment in the speaking and listening component' (Ofqual, 2013, p. 2).

Internationally, there is varied practice. For example, the Scottish Survey of Literacy includes an assessment of Listening and Talking using group discussion tasks at ages 8, 11 and 13. Oracy Australia (Education Department of Western Australia, 1997) offers oracy assessments for teachers to employ which focus on oral presentation, reading aloud, oral interpretation of literature and listening and responding. In the USA, the Common Core Standards for English Language Arts (CCSI, 2015), adopted by most states, provides a set of guidelines showing the expected standard for spoken language use at the end of each grade of schooling. However, none of these schemes include a framework which identifies the full range of skills required to meet the relevant assessment criteria.

2 The XXXXX Oracy Assessment Project

2.1 Background to the project

As mentioned earlier, the Educational Endowment Foundation, a UK-based charity, funded researchers at the University of XXXXX to work with School 21 on a two year project aimed to develop a curriculum for teaching oracy and an 'Oracy Assessment

Toolkit' for assessing students' levels of competence in oracy. The research team was to develop the toolkit for teachers to use with Year 7 students (age 11-12), enabling teachers' monitoring and assessment of student progress in oracy skills.

2.2 Method

The central aims of the research project were to create a Toolkit consisting of:

- an Oracy Skills Framework;
- a set of oracy assessment tasks;
- a rating scheme for assessing performance on the tasks and giving feedback to students.

In order to develop the skills framework we proceeded by examining existing frameworks and testing schemes and consulting with relevant experts in focus group sessions. Our expert panel consisted of eight members with a variety of expertise (see Acknowledgements section for a list of names and affiliations).

In the development of the tasks we trialed initial tasks with year 7 teachers and students near the start of the school year and end-of-year tasks later in the year with the same sets of teachers and students. The assessment rating schemes were trialed alongside the tasks. Throughout this trialing process we were revising our draft Skills Framework based on outcomes of the trials and further focus groups with our expert panel. Our consultative conference on Oracy Assessment in XXXXX, in which experts and practitioners were involved, also aided our development of the framework.

We also assisted School 21 staff in their aim of assessing the effects of their oracy-led curriculum by using the Toolkit to compare the performances of a sample of Year 7 children following the 'oracy-led' School 21 curriculum and a comparison school involved in the project (which we call MVC), which followed the usual National Curriculum (in which oracy is given relatively little attention).

The following sections address the key aspects of this process.

2.2.1 Developing the Oracy Skills Framework

As described above, most previous approaches to assessing oracy have relied on performance criteria related to specific situations, such as public speaking or group work. We planned a more general framework that represented the range of skills which could be drawn upon in any situation, enabling teachers to build an 'oracy profile' for any student which would not be situation specific.

In iteratively constructing the Oracy Skills Framework, we were influenced by theoretical conceptions of language use such as Hymes' ethnography of communication (Hymes, 1977) and the systemic functional linguistics of Halliday and his associates (Halliday, 1978). Much of such discussion has been concerned with second language acquisition

(e.g. Cummins, 1980; McNamara, 1997, Housen, Kuiken, & Vedder, 2012) but is relevant nevertheless. Building on earlier work by a range of applied linguists, Celce-Murcia, Dornyei & Thurrell (1995) offered what they called a 'pedagogically motivated model' of communicative competence designed for second language education which includes five components: (1) discourse competence; (2) linguistic competence; (3) actional competence; (4) sociocultural competence; and (5) strategic competence. While its breadth and subtlety are positive features, a disadvantage of that model, in our view, is that it seems to confuse the cognitive foundations of speech performance with observable features of talk and interaction. We wanted to create a framework that directed teachers' attention to what students actually said and did. In aiming to create a framework which would match not only expert understanding of the dimensions of competent language use, but also the concerns and perceptions of practitioners, we engaged in a series of consultations and discussions.

Discussions with our research partners in School 21 enabled sharing of professional and researcher expertise about what constitutes the effective use of spoken language and what might realistically be expected of 11 year olds in that respect when faced with the assessment tasks. Previously developed assessment tools for oracy were appraised. These included the APU assessments mentioned above and tools for assessing children's developing use of English as a second language. We reviewed debates about the importance of assessing complexity, accuracy and fluency (CAF: Housen & Kuiken, 2009) and tests such as IELTS (the International English Language Testing System) with its categories of fluency and coherence, lexical resource, grammar and pronunciation (IELTS, 2013). From these discussions and considerations the initial organising areas of the Skills Framework were devised.

Consultations were also undertaken with members of our expert group (see acknowledgements) - people of recognised stature in drama, English studies, sociolinguistics, applied linguistics and educational assessment. This group met three times during the development phase of the project (lasting approximately 14 months) and members were consulted individually and collectively by phone/email. We consulted teachers and gained the views of speech therapists, test developers and representatives of relevant organizations (including Cambridge Assessment, The Communication Trust, The National Literacy Trust, The Scottish Qualifications Authority, and the United Kingdom Literacy Association). Constructive criticism from such professionals enabled revision of both the framework and the assessment tasks. Reassuringly, however, the basic concept of a generalised skills framework for oracy and the chosen areas of the framework were supported by all.

Initially, the framework had eight main categories; but as a result of the iterative discussions with relevant experts as described above, this was reduced to four. The final version of the framework is presented in Figure 1. The four areas – physical; linguistic; cognitive; social and emotional - represent the different types of skill that are involved in the effective use of spoken language. The need to balance accuracy and complexity with clarity and practical usefulness means that the framework is presented in language unlikely to satisfy the rigorous criteria of an academic linguist. However, the intention

was to create a framework comprehensible to, and useable by, classroom teachers.

Insert Figure 1: Oracy Skills Framework

The Skills Framework defines the conception of 'oracy' that forms the basis of the accompanying assessment tasks (used to elicit evidence of these skills) and the assessment rating schemes (used to evaluate this evidence). Each of the four skill areas contain a number of specific skills that are listed on the left hand side. These are then described in detail on the right hand side of the diagram. More details and a glossary of skills can be found at:
XXXXX/ [2]

2.3 Task development

On arrival in secondary school, students will vary in their oracy skills, partly dependent on the extent to which their prior school and home experience has helped develop such skills. Tasks were therefore devised to allow teachers to make initial assessments together with a matched set of tasks for assessing progress at the end of students' first year. Validity was a major concern, and this meant ensuring that the tasks allowed the students to do 'what we want them to show us they can do' (Ahmed & Pollitt, 2011, p. 25). A set of Assessment for Learning (AfL) tasks were also devised which could be used throughout the year and adapted by the teacher to monitor development. The initial and end tasks were designed to sample a representative range of skills from the framework and, though there is some overlap in the skills assessed in each task, each one has a different emphasis. We were also concerned not to provide teachers with a large battery of tasks which it would seem impractical to achieve. That is, we aimed to generate a set of the minimum number of tasks which would best cover a wide range of skills. The tasks, and associated rating schemes, were trialled and refined against these criteria. Interviews were undertaken with both teachers and students, and adjustments were made to the final versions of tasks and assessment materials.

The first of the initial tasks devised was called 'Map', based on a task of the same name used by the Assessment of Performance Unit (APU, 1988). In this task, one student was given a map of a 'Treasure Island', and asked to plan a route from a port to a pirate's treasure hoard. However, they were told that this map was now out of date, and so they would need to consult a second student 'by phone' (in fact they sat back-to-back) who had an up-to-date version of the map. The task thus invoked such skills in our framework such as 'seeking information and clarification through questions' and 'taking account of level of understanding of audience'. Whilst trials confirmed the value of this kind of task for assessing communication skills, this particular task proved problematic in its original form. There was some confusion amongst students carrying out the task about what 'real

[2] Please note that all elements of the Oracy Assessment Toolkit - which includes the glossary, all assessment tasks and the rating scheme sheets - are available at:
www.educ.cam.ac.uk/research/projects/oracytoolkit/
As this is the case, they will not be included as appendices in this paper.

world' activity it was meant to simulate; and it did not encourage both students to take the role of explainer and questioner. Thus, an alternative paired instructional task, described below, was created for the end of year assessments.

The second initial task was a 'Talking Points' activity (Dawes, 2012). Talking Points are a set of somewhat controversial statements about a topic which students are asked to consider together and decide if, and why, they agree or disagree with the view expressed. This type of task has been found to be very effective for generating lively discussion (Mercer, Dawes & Staarman, 2009; Dawes, Dore, Loxley & Nichols, 2010). The aim of this task, which lasts ten minutes, was to get the students to use skills in managing a discussion, giving and seeking views supported by reasons, building upon each other's ideas and working towards consensus.

The third initial task was a Presentation task in which the students had to give a two-minute presentation to camera. They were given some preparation time with their teachers before the task; they were not allowed to use a script, but could bring a prompt card with them. Each student was allowed an unrecorded trial run and then they presented the 'real thing' to camera.

The three end-of-year tasks were designed to be parallel forms of the initial tasks, so that comparisons could be made between student performances. The skills involved in each task were the same for the initial and end of year versions, enabling teachers to make an assessment of the children's progress in oracy. To replace 'Map', a task based on Lego construction materials was created which allowed for assessment of the same skills. In this task, one student had a picture of a completed Lego model, while their partner had a box of Lego parts (which included not only those required for the model but also several others). They were asked to work together, sitting back-to-back, to enable the second student to build the model. Importantly, by switching roles, this new task enabled each student to take the role of both builder and guide.
As well as the three formal assessment tasks, five Assessment for Learning (AfL) tasks were devised for teachers to use in a whole class context. They focused on debate, drama, role play, group talk and presentation. Each can be adapted to suit a teacher's choice of content; teachers are provided with loose guidelines in which the talk objectives for the task are listed, followed by examples and assessment procedure guidelines. As well as teacher assessment, these tasks involve students' self and peer assessment (Clarke, 2001).

2.4 Rating scheme development

A three way rating scheme was developed for teachers to judge a student's performance on each skill for each task. This used a version of a mastery model in which students are judged as demonstrating each skill consistently, only some of the time, or not at all. With such tasks, responses can be rated by making either holistic or analytic judgements. Holistic judgements involve giving the performance as a whole a single grade. Analytic judgements involve giving a series of grades for different aspects of the performance and then aggregating these, either with or without weightings. Harsch and Martin (2013) found no agreement amongst researchers on whether holistic or analytic methods yield

more reliable and valid scores of students' writing; no comparable work has been done for oracy. In our scheme, assessors first make a rating based on a subset of skills from the Skills Framework, and then make an overall rating. This represents a combination of analytical and holistic approaches.

To ensure validity, focused marking criteria highlight the key skills involved in each task, as represented in the Skills Framework. Performance descriptors were based on observed characteristics of different levels of performance (Greatorex, Johnson & Frame, 2001; Fulcher, Davidson & Kemp, 2011). A panel of eight teachers from School 21 and MVC was set up. The research team led discussions with these teachers, aimed at bringing them to a common view of relevant skills; and we expected this to influence their judgement of students. Teachers were given an assessment sheet with a list of the relevant skills and a space to rate performance on each skill. They could rate a performance as Gold ('consistently demonstrates this skill'), Silver ('demonstrates this skill some of the time') or Bronze ('rarely or never demonstrates this skill yet'). During the development phase there were no exemplars for teachers to use to help them to judge the standard. As this was a research and development project, the devising of marking criteria – essentially what distinguishes a Gold performance from a Silver or Bronze – was central to the way in which the project was conceived. It was by engaging professionals in discussions around student responses to the tasks, and around the definition of oracy framework skills, that we were able to structure the video/descriptor items (accessed through our website) that form the guide materials for the oracy assessment toolkit. As a result of this work, in the final version of the toolkit there are exemplar videos of students performing at different levels, along with descriptors of the oracy skills seen in these exemplars. These help to benchmark the standard of performance on each of the tasks. During the trials it became apparent that some teachers wanted to use a more fine-grained rating scheme and were using Bronze+, Silver + and Gold + to distinguish more levels of performance. We therefore added these finer levels of assessment to the assessment sheets used in the subsequent phase and went further in this refinement as the work with teachers progressed, as described below.

3 Trialling the tasks and assessment procedures

3.1 Schools and sample

Initial versions of the tasks were trialled in four schools. School 21, our primary partner in this project, mainly takes students from a multicultural population of low socio-economic status in London; and it has an 'oracy-led' curriculum. The other main school involved, MVC is a rural state comprehensive (i.e., non-selective) secondary school in eastern England that works to the National Curriculum, with a predominantly middle class intake. We also trialled early versions of the tasks in two triangulating schools, producing videos that were used with School 21 and MVC teachers to reduce the risk of bias in their observations. CWS is an urban state comprehensive secondary school in central England with a predominantly working class population of varied ethnic backgrounds; and CS is a comprehensive secondary school in a small market town in the north of England, serving a socially mixed but largely white indigenous population.

Given that our aim was to create an assessment tool that could be used in any mainstream school, the natural variation amongst these four schools enabled us to test the assessment tasks and rating scheme on a suitably diverse population. Of these latter three, only CWS had previously been involved with the research team, through the provision of professional development sessions for its teachers on 'developing language for learning'.

Because of the geographical distances involved and the timetable of the project, it was only possible to include students (and teachers) from School 21 and MVC in the trialling of the final versions of the tasks. Students were selected from those whose parents had given permission for their involvement (only 2 students were excluded for such reasons). The researchers asked teachers to select a range of students who, on the basis of their initial impressions, represented a range of competence levels in spoken English.

In School 21 there were seven boys and five girls in the focus group, with reading ages ranging from 7.1 to 15+[3] and with a mean of 11.0. At MVC there were six boys and six girls in the focus group, with reading ages ranging from 8.6 to 14+ and with a mean of 12.0.

3.2 Procedures for developing the Assessment Toolkit

Sessions were video-recorded to enable the close analysis of students' performances, to facilitate discussion with teachers and others for standard-setting purposes, and to provide exemplars on the Oracy Toolkit website. Two teachers in each school were asked to rate the performance of each student 'live' in response to a range of skills most pertinent to that task.

The development of the assessment criteria and scoring scheme also involved a review day with five teachers involved in the project. Videos of the performances of 16 students, from School 21, MVC and CS, were used as test materials. Using pairs of videos of students (selected by their teachers in terms of their availability and to provide some mixed gender pairings) carrying out each of the initial tasks, the teachers were asked to judge, for each of the tasks, which student was 'better' in terms of oracy. Allowing repeated viewing of videos when requested, new comparisons were then made between individual students. The same panel of teachers from School 21 and MVC was involved; and they ranked students from both schools. This meant that some teachers knew some of the students they were assessing, but not all. As mentioned earlier, through initial discussions with the research team we aimed to bring the teachers to a common view of relevant skills; and we expected this to influence their ranking of performances. During the ranking exercise the teachers could refer to the skills framework to focus their observations; however, they did not any material that indicated levels of performance, as the purpose of the work at this point was to refine, through discussion, what characterised performances at different levels..

The 5 teachers (DS, LG, GV, SH and AS in Table 2) were asked to select the 'top 2' students on each task; and then to make a paired comparison of these students (with video

[3] Reading ages above 14 are represented as whole years with a + sign in English schools.

viewing again allowed if requested). This iterative process continued until an overall rank order of all the children for each task was decided. A different set of eight students was involved in each set of comparisons. There was a high level of consensus in the initial paired comparisons for each task, recorded as presented in the example Table 1 below.

- *Insert Table 1. Presentation task initial pair scores (1 beats 0)*

The rank order of students by the teachers on this video review day matched the teachers' initial ratings for these performances for the Presentation task. For the Map task there is no contradiction in the rank order but most of these performances were rated 'gold' overall so there was a lack of discrimination in this sample. However, the teachers were able to discriminate amongst the Gold performances when reaching an agreement on rank order. For the Talking Points task the rank order was less consistent with earlier ratings. The anomalies here were associated with the use of assessments from one of the project schools, where the teacher had not been trained in the use of the assessment toolkit and where a high number of Gold assessments had been made. Similar review days were carried out with the expert panel for the project and with a panel of five secondary school English teachers who were not involved in the project. At each, the participants rated students' responses, discussed both their ratings and the relevance of particular skills within the assessments, and commented on all aspects of the tasks.

Participating students (in all 4 schools) were interviewed in pairs or groups of three about their experiences of the tasks, using open questions as a stimulus for discussion. Their views were sought on the tasks, focusing on the clarity of instructions, the level of difficulty and perceived value in assessing their oracy skills. In School 21 and MVC, after the three initial tasks had been trialled and again after the three end of year tasks had been trialled, the teachers were interviewed. The interviews and the review day discussions led to significant developments in the tasks for the final version of the Toolkit, most noticeably the replacement of the Map task by the Lego task. Typical comments here were:

> "…The new map and the old map was hard because I didn't know. I thought that because it was a new map I hadn't thought that it was the same map but just new things on it, but it was different roads." (Student interview)

> "they weren't really empathising with what their colleague could see … I think the map task locked some of them out, actually. I think it was the least appealing to them and the one they found the hardest to do." (Teacher comment)

Although some of the students found the replacement Lego task hard, they found it more comprehensible and purposeful than the Map task.

> "I found the Lego easier because someone was telling the other person where everything was. Then, with the map it wasn't very clear, because they weren't the same. So, it was quite hard to know where everything was." (Student interview)

And the teachers and experts agreed that the Lego task was an improvement:

> "when we looked at the Lego task in comparison with the map task, they were able to [complete the] task very much better." (Teacher comment)

> "The task is genuine; you're not pretending that you're sitting making a Lego model, you actually are making a Lego model and somebody else has got a picture of it." (Expert comment)

The expert panel felt that the Talking Points task allowed students to demonstrate the relevant skills identified, though they considered that some of the Talking Points about 'talk' were not controversial enough to provoke a lively debate and thus for students to demonstrate relevant skills. This was addressed in the end version of this task by involving the originator of the Talking Points activity format (Dawes, 2012) in a team session to devise new sets.

For the Presentation task, interview responses indicated that **t**he students saw how useful the skills involved in this task would be. For example, one commented:

> "I think it is because like when we get older we're going to have to like speak to people like face to face that we haven't met before and if you…and in front of cameras like for a job interview we might be like that." (Student interview)

A general issue raised by both the experts and teachers was that of the clarity of the task instructions given to students and so these were modified accordingly. The interviews and video review days were a vital aspect of the work, allowing both theoretical and professional perspectives to be taken into account in the toolkit design.

4 Results: assessing oracy with the Toolkit

4.1 Skills ratings on initial and end tasks

The initial tasks had been administered in School 21 and MVC in September, during the first part of the Autumn term. The end tasks were used in both schools in March/April. The final calibration of the assessment scheme involved teachers from School 21 and MVC. (Teachers from CS and CWS were invited, but attendance proved impossible.) Teachers used Bronze, Silver, Gold to make their ratings, and some also used Bronze+, Silver -, Silver + and Gold -. For analysis the Bronze, Silver and Gold ratings were converted into numbers:

Bronze	1
Bronze +	2
Silver -	3
Silver	4
Silver +	5
Gold -	6

Gold 7

The teachers' mean overall task ratings for School 21 and MVC, the 'control' school, can be seen in Tables 2-4.

- *Insert Table 2: Individual Presentation Task mean ratings – teacher assessments*

- *Insert Table 3: Group Talking Points Task mean ratings – teacher assessments*

- *Insert Table 4: Paired instructional Task mean ratings – teacher assessments*

The differences between initial and end task means can be seen as a measure of progress. Limited progress can be seen in these ratings for School 21 in the individual presentation and group discussion tasks, but not in the paired instructional task. This may have been because of the discrepancies between the Map and the Lego tasks. (The only initial test scores available for comparison were of course based on the Map task.) The mean ratings for MVC show little progress. Teacher and expert discussions noted that the low progress measures may be explained by high initial task ratings by the teachers, who knew the students (albeit by only a few weeks of the school term) they were rating by the time of the initial assessments.

Due to the possible bias in the data caused by teachers knowing students (as noted above), the progress of School 21 and MVC students was also assessed using only researcher ratings. There is, of course, the potential for bias derived from the researcher knowledge of the data, but here it should be noted that the researchers were only familiar with the students through their task responses; they did not have the holistic view of the students that may have influenced the assessments from the teachers, who worked with the students regularly in their schools in a range of contexts.

The three researchers first rated students independently, and then through discussion sought a consensus score, as recorded in Tables 5-7. The small sample in each school prevents any differences being judged as statistically significant.

- *Insert Table 5: Individual presentation task mean ratings- researcher assessments*

Researcher assessments for School 21 show clear progress in the oracy skills required for the Presentation task, with a difference of 1.64 between initial and final mean ratings. For MVC there is a very slight trend towards improvement but no real difference in the mean ratings.

- *Insert Table 6: Group 'Talking points' task mean ratings- researcher assessments*

In the Talking Points task for School 21, there is again an indication of progress in the oracy skills required, with a progress rating of 1.08. For MVC there is a trend in a positive direction but the progress rating is less than that for School 21.

- *Insert Table 7: Paired instructional task mean ratings – researcher ratings*

In the instructional (Map/Lego) task, for School 21 there is again an indication of progress in the oracy skills required but the difference in means is very small at 0.58. For MVC there is again a slight trend in a positive direction.

The instructional task data in particular must be interpreted with caution since, as explained earlier, the replacement of Map with Lego means that differences in performance on this task may have been an artefact of these changes. However, it can be seen that School 21 students still achieved noticeably higher ratings than those in MVC on both initial and end tasks.

Overall, the ratings for the School 21 students were higher than those of the MVC students on two initial tasks and all three follow-up tasks. The higher initial scores for School 21 may be because of factors outside our control, by the time it was possible to administer the initial tasks (in October), the School 21 students had already begun studying the schools' distinctive 'oracy-led' curriculum, which provided training in both presentational and group interaction skills.

4.2 Comparative ratings and reliability

Video review sessions with researchers, project teachers, the project expert panel and a new panel of independent teachers were used to test the reliability of the rating scale. An established method for averaging correlations (Hatch & Lazaraton, 1991) was employed for calculating inter-rater reliability (IRR), as the more traditional Kappa statistic was unsuitable for our data given the small number of students and lack of variability within judges' ratings. The first step was to generate a Pearson correlation matrix and then calculate the average correlation after a Fisher Z transformation (to transform to a normal distribution and correct for the fact that these are ordinal data). The derived average of the transformed correlation coefficients, r_{ab} was then substituted using the formula:

$$r_{tt} = \frac{n.r_{ab}}{1+(n-1)r_{ab}}$$

r_{tt} = reliability of all the ratings
n = number of raters
r_{ab} = correlation between two raters (or average correlation if there are more than two).

The reliability of all the ratings r_{tt} was then transformed back to a Pearson correlation.

Table 8 gives the IRR values for each of the three initial tasks and three end-of-year tasks

for School 21 and MVC. The shared percentages of the variance are also given for each of the tasks. This is the portion of variance in the data that represents the shared consensus of the judges, with the rest of the variance being due to lack of perfect reliability in using the rating scales and inconsistencies in how they interpreted the students' performances.

- *Insert Table 8: IRR values for each of the three initial tasks and three end-of-year tasks for School 21 and MVC*

** This result is due to the lack of variability within one judge's ratings for this task, leading to an unreliable calculation of the correlation.*
*** The cells marked 'Incomplete' are where IRR could not be calculated due to missing data for each of the judges.*

To help ensure future reliability in the use of the Toolkit, a library of video exemplars has been provided on the project website to give teachers a benchmark standard of performance for each level. The examples have empirically derived level descriptors and should help teachers to rate their own students' performances in a reliable manner and give specific and informative feedback to the students on how well they demonstrate the various oracy skills in different contexts.

5. Discussion and conclusions

The XXXXX Oracy Assessment Project's main aim was to produce an Assessment Toolkit that combined research-based validity with a practical ease of use for teachers. Evaluations by our expert panels and feedback from participating teachers encourage us to believe that, to a reasonable extent, this aim has been achieved. Feedback from both teachers and students in the participating schools suggests that they perceived the tasks as valid tests of communicative skill – though several students commented that being video-recorded and observed by researchers made them more nervous and less fluent than they thought they would be in a more private situation. One can only hope that the use of the Toolkit in more normal school circumstances as intended, would reduce this problem. In the final version of the assessment tasks, instructions have been made clearer; allowances have been made for the use of a finer grained scale with + and – as well as Bronze, Silver and Gold; and suggestions have been made in the Toolkit instructions about how to use assessments to design suitable teaching activities. We concede that the tasks might still be improved, given more time and resources; but they have been developed with due care and are now available for public scrutiny and use. The Oracy Assessment Framework has had a very positive response from experts and teachers, though we are very aware that it may lack subtlety and detail in the eyes of applied linguists.

In Section 1.3 we discussed the need for oracy assessments to take into account the social situation in which talk is used, and the risks of reducing an interactive social phenomenon to a series of set tasks. A criticism of the Toolkit might be that the presentation, group work and problem solving activities are not 'real' in that they are set by researchers/teachers and do not naturally arise in the lives of the participating students.

There is also the possibility that nervousness generated by a test situation disrupts a student's normal performance. However, similar criticisms can be made of formal tests of reading, numeracy, intelligence and so on. To the extent that the students recognise the importance of oracy skills (which interview data suggested our participating students did) and that tasks are clear and comprehensible in their procedures and criteria for success (which feedback suggests ours are), then one can expect that most will be motivated and able to perform to the best of their abilities. In other words, these issues can be dealt with just as effectively in relation to the assessment of oracy as they can for the assessment of other comparable skill sets.

One aspect of the overall project which we cannot deal properly with in this article is the evaluation of the innovative 'oracy-led' curriculum introduced by School 21. While small sample sizes must limit the confidence of claims, the higher scores of the School 21 students suggest that the development of students' spoken language skills is aided by their teachers (a) involving them in awareness-raising activities which encourage students' metacognition about ways of talking (as discussed in Mercer, 2013); (b) providing them with instruction on how to use talk effectively in different circumstances; and (c) embedding spoken language practice into the curriculum for all subjects (not just English, drama or modern languages). The higher skills scores were the major difference across the schools; but it was also interesting to note that our qualitative analyses of the video data showed that School 21 students: (i) took longer turns which allowed them to provide clearer explanations and reasons in group tasks (items 5, 7a and 10a in the Skills Framework, Figure 1); (ii) organised group tasks more effectively (items 12a and b); and, particularly in the presentation task, were (iii) more fluent and confident in their overall performance (items 1 and 14a). The value of an 'oracy-led' curriculum as used in School 21 has thus been supported by use of the Oracy Toolkit.

The outcomes of the project show that it is possible to provide teachers with (a) a framework for understanding the spoken language skills that their students will need to use talk effectively in the various social situations they find themselves in; (b) a set of tasks for assessing their students' oracy skills across a sample of such situations; and (c) a rating scheme which provides a valid and fairly reliable way of assessing individual students' levels of competence and the progress they make over time. Part of the motivation in developing the Toolkit was to raise the status of oracy as a crucial set of life skills on a par with those of literacy and numeracy. Once they leave school, most young people will find that skills in using spoken language will be required much more often than those of anything but the most basic skills of numeracy. By showing how spoken language skills can be monitored and assessed, and by identifying the skills that an 'oracy-led' curriculum can help to develop, we hope to help oracy achieve the place in school curricula for the 21st century that it deserves.

References

Ahmed, A. & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in education: Principles, policy and practice*, *17*, 259-278.

Alexander, R. (2012). *Improving oracy and classroom talk in English schools: Achievements and challenges.* Available at: http://www.primaryreview.org.uk/downloads_/news/2012/02/2012_02_20DfE_oracy_Alexander.pdf [Accessed 15 October 2013].

APU (1988). *Language performance in schools: Review of APU language monitoring 1979-1983.* London: Her Majesty's Stationary Office.

Barnes, D. (1980). Situated speech strategies: aspects of the monitoring of oracy. *Educational review*, *32*, 123-131.

Brooks, G. (1989). The value and purpose of oracy assessment. *English in Education*, *23*, 87-93.

CCSI (2015). *English Arts Standards*. Common Core Standards Initative. Available at: http://www.corestandards.org/ELA-Literacy/ [Accessed 7 May 2015].

Celce-Murcia, M., Dornyei, Z., & Thurrell, S. (1995). Communicative Competence: a pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6*, 5-35.

Cinamon, D. & Elding, S. (1998). Tracking Talk. In J. Holderness & B. Lalljee (Eds.), *An introduction to Oracy: Frameworks for talk* (pp. 212-233). London: Cassell.

Clarke, S. (2001). *Unlocking formative assessment*. London: Hodder and Stoughton.

Communication Trust (2013). *Speech, language and communication progression tools*. Available at: https://www.thecommunicationtrust.org.uk/resources/resources/resources-for-practitioners/progression-tools-secondary/practitioners/universally-speaking/ [Accessed 10 April 2014].

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: implications for bilingual education and the optimal age issue. *TESOL Quarterly, 14*, 175-187.

Dawes, L. (2008). Encouraging students' contributions to dialogue during science. *School Science Review*, *90*, 1–8.

Dawes, L. (2012). *Talking points: Discussion activities in the primary classroom*. London: David Fulton/Routledge.

Dawes, L., Dore, B., Loxley, P. & Nichols, L. (2010). A talk focus for promoting enjoyment and developing understanding in science. *English Teaching: Practice and*

Critique, *9*, 99-110.

DfES (2003). *Speaking, listening, learning: Working with children in key stages 1 and 2.* London: Department for Education and Skills.

Education Department of Western Australia (1997). *Oral language resource book.* Rigby: Heinemann.

Fulcher, G., Davidson, F. & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*, 5-29.

Goswami, U. (2009). Mind, brain, and literacy- biomarkers as usable knowledge for education. *Mind, Brain, and Education, 3,* 176-184.

Goswami, U. & Bryant, P. (2007). Children's cognitive development and learning. *Research Report 2/1a: The Primary Review.* Cambridge: University of Cambridge.

Greatorex, J., Johnson, C. & Frame K. (2001). Making the grade: developing grade descriptors for accounting using a discriminator model of performance. *Westminster Studies in Education, 24,* 167-181

Halliday, M.A.K. (1978). *Language as social semiotic.* London: Edward Arnold.

Harsch, C. & Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assessment in education: Principles, policy & practice, 20,* 281-307.

Hart, B. & Risley, T.R. (1995). *Meaningful differences in the everyday experience of young American children.* New York: Brookes.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* New York: Newbury House.

Housen, A. & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30,* 461–473.

Housen, A., Kuiken, F. & Vedder , I. (Eds.) (2012) *Dimensions of L2 performance and proficiency*. Amsterdam: John Benjamins.

Howe, A. (1991). *Making talk work. NATE papers in education*. London: National Association for the Teaching of English.

Howe, C., & Abedin, M. (2013). Classroom dialogue: a systematic review across four decades of research. *Cambridge Journal of Education, 43*, 325–356.

Hymes, D. (1977) *Foundations in sociolinguistics.* London: Tavistock.

IELTS (2013) *Assessment Criteria: Speaking*. Available at: http://www.ieltsessentials.com/pdf/BandcoreDescriptorsSpeaking.pdf [Accessed 5 November 2013].

Latham, D. (2005). Speaking, listening and learning: a rationale for the Speaking and Listening Profile. *English in Education, 39*, 60-74.

Maybin, J. (1988). A critical review of the DES Assessment of Performance Unit's Oracy Surveys. *English in Education, 22,* 3-18.

McNamara, T. (1997). 'Interaction' in second language performance assessment: whose performance? *Applied Linguistics, 18,* 446-466.

Mercer, N. (2008). Talk and the development of reasoning and understanding. *Human Development, 51*, 90-100.

Mercer, N. (2013). The Social Brain, language, and goal-directed collective thinking: a social conception of cognition and its implications for understanding how we think, teach and learn. *Educational Psychologist, 48,* 148-168.

Mercer, N., Dawes, L. & Staarman, J.K. (2009). Dialogic teaching in the primary science Classroom. *Language and Education, 23,* 353-369.

Monroe, S. (2009) *Discover your voice. Debating resources for Key Stage 2*. London: The English-Speaking Union Centre for Speech and Debate

Norman, K. (Ed.) (1992). *Thinking voices: The work of the National Oracy Project*. London: Hodder & Stoughton.

Ofqual (2013). Analysis of responses to the consultation on the proposal to remove speaking and listening assessment from the GCSE English and GCSE English language grade. Ofqual Report number 13/5317. Available at: http://dera.ioe.ac.uk/17586/7/2013-08-29-analysis-of-responses-to-the-consultation-removal-of-speaking-and-listening.pdf [Accessed 7 January 2014]

Oliver, R., Haig, Y. & Rochecouste, J. (2005). Communicative competence in oral language assessment. *Language and Education, 19,* 212-222.

Pinker, S. (2007) *The stuff of thought: Language as a window into human nature*. London: Penguin.

Rojas-Drummond, S., Littleton, K., Hernández, F. & Zúniga, M. (2010). Dialogical interactions among peers in collaborative writing contexts. In K. Littleton and C. Howe (Eds.) *Educational dialogues: Understanding and promoting productive interaction* (pp. 128-148). Abingdon: Routledge.

Van den Bergh, H. (1987). *Large scale oracy assessment in the Netherlands.* Research and technical report 143. Amsterdam: SCO.

van der Wilt, F., van Kruistum, C., van der Veen, C., & van Oers, B. (in press). Gender differences in the relationship between oral communicative competence and peer rejection: An explorative study in early childhood education. *European Early Childhood Education Research Journal*, 25(2)

Van Oers, B., Elbers, E., van der Veer, R. & Wardekker, W. (Eds.) (2008). *The transformation of learning: Advances in cultural-historical activity theory*. Cambridge: Cambridge University Press.

Vass, E. & Littleton, K. (2010) Peer collaboration and learning in the classroom. In: K. Littleton, C. Wood & J. K. Staarman (Eds.) *International handbook of psychology in education* (pp. 105-136). Leeds: Emerald .

Vygotsky, L. (1962). *Thought and language*. Cambridge MA: MIT.

Wilkinson, A. (1965). *Spoken English.* Edgbaston, Birmingham: University of Birmingham.

Wilson, M., Neja, I., Scalise, K., Templin, J., Wiliam, D., Torres Irribarra, D. (2012). perspectives on methodological issues. In P. Griffin, B. McGraw and E.Care, (Eds.) *Assessment and teaching of 21st century skills* (pp. 67-141). New York: Springer.

Whitebread, D., Mercer, N., Howe, C. and Tolmie, A. (Eds.) (2013). Self-regulation and dialogue in primary classrooms. *British Journal of Educational Psychology Monograph Series II: Psychological Aspects of Education: Current Trends, 10*. Leicester: British Psychological Society.

Wright, J., Brinkley, I. & Clayton, N. (2010). *Employability and skills in the UK: Redefining the debate*. London: The Work Foundation.

Acknowledgements

We would like to acknowledge the positive collaborative efforts of our partners in this project, School 21; the support of funders the Education Endowment Foundation and project evaluators Sheffield Hallam University; the teachers and students in other schools that took part in this project - Melbourn Village College, Cambridgeshire (MVC), Cardinal Wiseman School and Language College, Coventry (CWS) and Cockermouth School, Cumbria (CS); and the project expert advisory panel:

Alan Howe, Formerly of the National Strategies
Greg Brooks, Professor of Education, University of Sheffield
Janet White, JW English Consultancy Ltd
Adrian Beard, Assessment and Qualifications Alliance
Lyn Dawes, Educational Consultant
Stephanie Merritt, Speech Therapist
Lesley Hendy, Voice educator
Evelina Galaczi, Cambridge English

Table 1

Table 1. Presentation task initial pair scores (1 beats 0)

Teacher	DS	LG	GV	SH	AS
Student 1	0	0	0	0	0
Student 2	1	1	1	1	1
Student 3	0	0	0	0	0
Student 4	1	1	1	1	1
Student 5	0	0	0	0	0
Student 6	1	1	1	1	1
Student 7	0	0	0	0	0
Student 8	1	1	1	1	1

Table 2

Table 2: Individual Presentation Task mean ratings – teacher assessments

	School 21	MVC
Initial task mean	3.64	4.46
End task mean	4.11	4.23
Difference in means	0.47	0.23

Table 3

Table 3: Group Talking Points Task mean ratings – teacher assessments

	School 21	MVC
Initial task mean	4.20	4.50
End task mean	5.00	4.02
Difference in means	0.80	-0.48

Table 4

Table 4: Paired instructional Task mean ratings – teacher assessments

	School 21	MVC
Initial task mean	4.61	4.75
End task mean	3.83	5.00
Difference in means	-0.78	0.25

Table 5

Table 5: Individual presentation task mean ratings- researcher assessments

	School 21	MVC
Initial task mean	2.91	3.90
End task mean	4.55	4.00
Difference in means	1.64	0.10

Table 6

Table 6: Group 'Talking points' task mean ratings- researcher assessments

	School 21	MVC
Initial task mean	4.17	2.33
End task mean	5.25	3.00
Difference in means	1.08	0.67

Table 7

Table 7: Paired instructional task mean ratings – researcher ratings

	School 21	MVC
Initial task mean	4.17	2.25
End task mean	4.75	2.50
Difference in means	0.58	0.25

Table 8

Table 8: IRR values for each of the three initial tasks and three end-of-year tasks for School 21 and MVC

Task	**IRR**	**Shared Percentage of Variance**
S21Map	0.69	48%
S21Lego	0.73	54%
S21 Pres1	0.28 *	8%
S21 Pres2	0.77	59%
S21 TP1	0.64	41%
S21 TP2	0.62	38%
MVC Map	Incomplete **	Incomplete
MVC Lego	0.72	52%
MVC Pres1	0.88	77%
MVC Pres2	0.90	81%
MVC TP1	Incomplete	Incomplete
MVC TP2	0.83	69%

** This result is due to the lack of variability within one judge's ratings for this task, leading to an unreliable calculation of the correlation.*
*** The cells marked 'Incomplete' are where IRR could not be calculated due to missing data for each of the judges.*

Figure 1

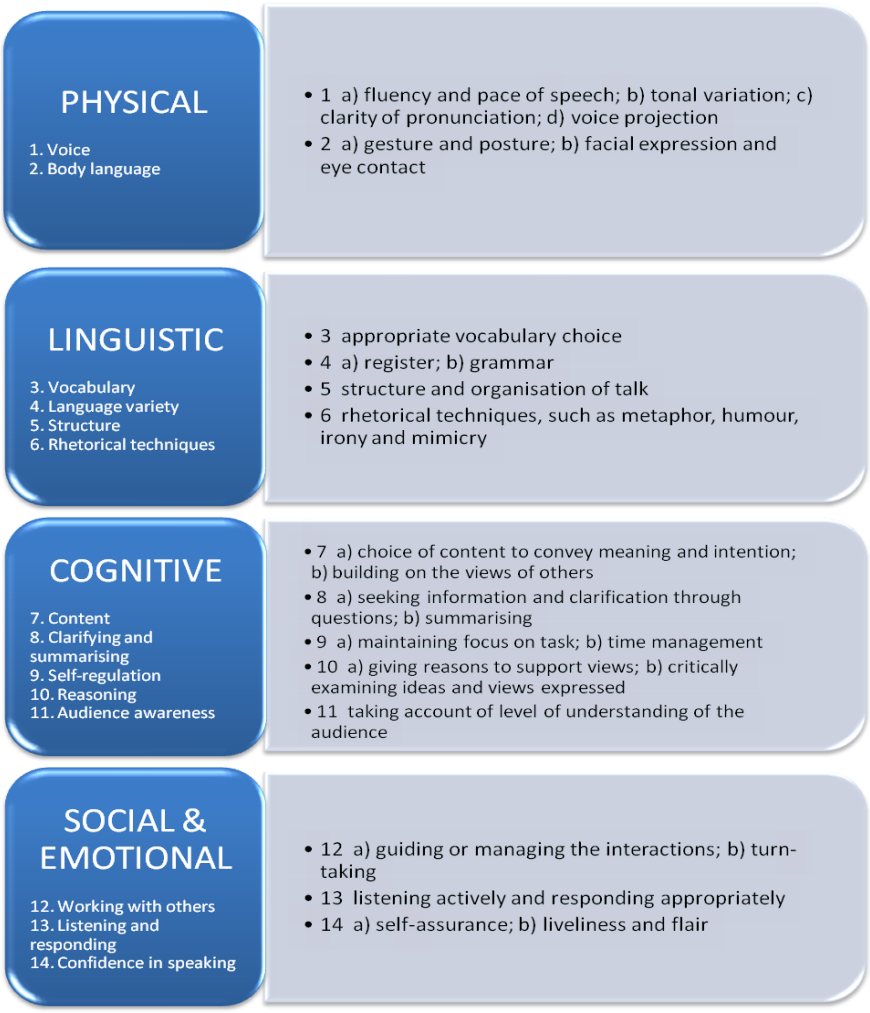

Figure 1: Oracy Skills Framework